Enhancing Mathematical Reasoning in Large Language Models through Reasoning Distillation, GRPO, and Multi-agent PRM Reranking

Abhinav Agarwal, Carlo Baronio, Shubhra Mishra, Shree Reddy

Stanford
Computer Science

Introduction

The AIMO Progress Prize challenge represents a significant milestone in advancing AI capabilities for mathematical reasoning. Success in this competition requires solving complex mathematical problems at the national olympiad level, demanding sophisticated reasoning, creativity, and precise computation. Existing LLMs struggle with consistency and accuracy in advanced mathematical reasoning.

Current Challenges in LLM Math Reasoning:

- Cold start issues in reinforcement learning for reasoning tasks.
- Robust verification and refinement of generated solutions.
- Maintaining consistent performance across diverse mathematical problem types under runtime constraints.

Background

- GRPO is a more efficient alternative to PPO. It samples a group of trajectories and normalizes its rewards, using the mean as the baseline instead of a value network.

$$\frac{1}{G}\sum_{i=1}^{G}\frac{1}{|o_i|}\sum_{t=1}^{|o_i|}\left\{\min\left[\frac{\pi_\theta(o_{i,t}|q,o_{i,<t})}{\pi_{\theta_{old}}(o_{i,t}|q,o_{i,<t})}\hat{A}_{i,t},\text{clip}\left(\frac{\pi_\theta(o_{i,t}|q,o_{i,<t})}{\pi_{\theta_{old}}(o_{i,t}|q,o_{i,<t})}\right.\right.\right.\left.\left.\left.-\beta\mathbb{D}_{KL}\left[\pi_\theta||\pi_{ref}\right]\right\}\right.$$

- Counterintuitively, with KL coefficient 0, a gradient step per batch, and advantage normalization, the **loss is always 0** but with non-zero gradients.

$$J(\theta)=\frac{1}{G}\sum_{i=1}^{G}\left(\frac{\pi_\theta(a_i|s)}{[\pi_\theta(a_i|s)]_\nabla}A_i-\beta D_{KL}(\pi_\theta\parallel\pi_{\text{ref}})\right)$$

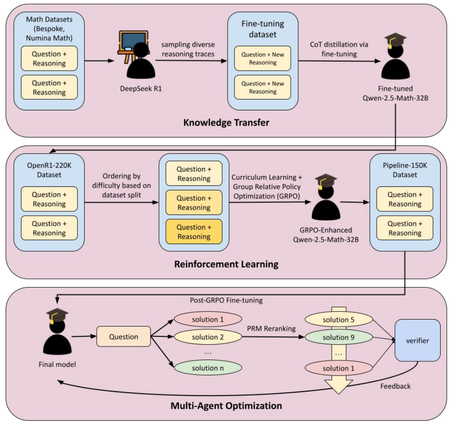
- Outcome reward function:

$$R=\begin{cases}0.2, & \text{if the output uses correct format}\\ 1.0, & \text{if the output is correct}\end{cases}$$

Data

Bespoke 17K	Problems, answers, reasoning traces by sampling R1
Numina Math 1.5	860K+ competition-level problems and solutions. **We created our own long CoT reasoning dataset of 150K problems using this dataset filtering out for OpenR1 and Bespoke.**
Open R1 220K	Built on top of NuminaMath, contains 2-4 reasoning traces for each problem

Methods



Reasoning Capability Distillation

- Use DeepSeek R1 as teacher; Qwen2.5-32B as student.
- Specialized Chain-of-Thought (CoT) distillation leveraging NuminaMath 1.5 dataset.
- Validate successful reasoning transfer on benchmark datasets to ensure robust baseline.

Group Relative Policy Optimization (GRPO)

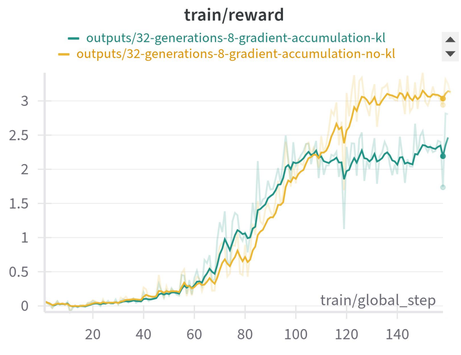
- Reinforcement learning on our finetuned Qwen2.5-32B model.
- Group-based advantage estimation reduces memory usage, curriculum learning gradually increases problem complexity.

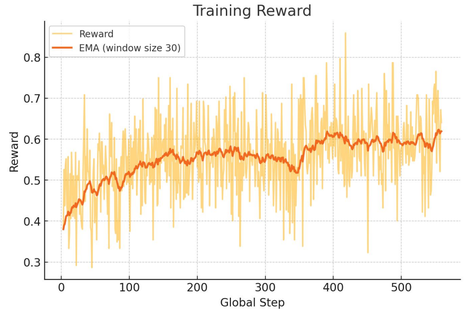
Post GRPO Tuning and Multi-agent PRM Reranking

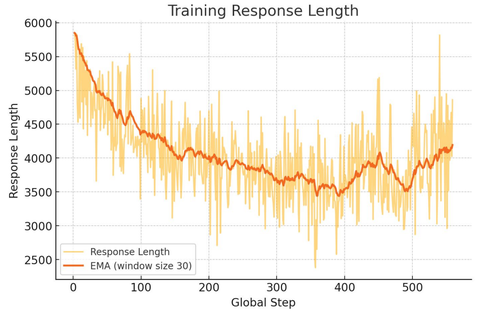
- Multiple independent inference attempts per problem.
- Preference Reward Models (PRMs) rerank and verify generated solutions.
- Multiple inference passes for solution refinement, ensemble methods used to aggregate best solutions.

GRPO Experiments & Observations

- KL penalty: we tested GRPO on Qwen2.5-1.5B and GSM8K, and found that a smaller KL coefficient improves the learning
- Moreover, if we the KL coefficient is 0, we can discard the reference model, saving both memory and compute
- By forking OpenRLHF, we are able to colocate inference and training on the same GPUs, offloading deepspeed/vLLM to CPU memory when needed
- We are able to scale up GRPO to DeepSeek-R1-Distill-Qwen-7B with a subset of OpenR1-Math-220k



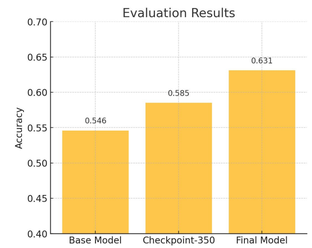


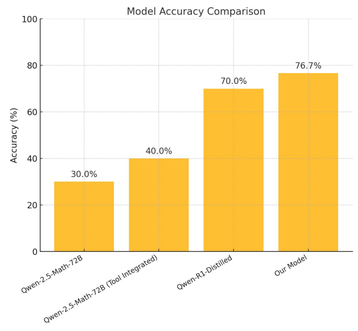


- Non-reasoning model, but distilled from R1 reasoning trajectories
- During GRPO the model learns how to use the CoT tokens more efficiently, and after that the response length of hard tasks starts to go up again -> extend context window

Results

- We tested our GRPO-trained model on a holdout dataset of 1k tasks from OpenR1-Math-220K
- The reasoning traces of the final checkpoint are shorter and achieve higher accuracy than the base model
- 2.1x faster inference





- We tested our final model on **AIME 24 benchmark**.
- The token limit for our model was set to 5048 compared to 32K tokens for the baseline and the R1-Distilled models.
- Our 32B model **achieved significantly higher pass@1 score** compared to both Qwen2.5-Math-72B and DeepSeek-R1-Distill-Qwen-32B while being much more token efficient.

Analysis & Conclusions

- Non-reasoning models distilled from R1 reasoning trajectories learn to efficiently utilize CoT tokens
- During GRPO, our model shows stable training with KL coefficient 0, allowing efficient scaling to 7B/32B models
- Key results show our approach yields:
  - **2.1x faster inference** while maintaining accuracy
  - Significant improvement on AIME'24 benchmark problems
  - **Higher pass@1** scores vs. comparable baselines
- The combination of distillation, GRPO, and multi-agent verification creates complementary benefits. Our approach efficiently handles context limitations through better reasoning compression

Future work: Extending mathematical reasoning capabilities to more complex domains and investigating token efficiency optimizations for longer problems. Also, training reasoning models on CoD (Chain of Draft) data.

Code: github.com/abhinav30219/math_reasoner

Model: huggingface.co/abhinav302019/Math-Reasoning-Model-32B